DOCUMENT RESUME

ED 174 669 TM 009 661

AUTHOR Shapiro, Jonathan
TITLE Assessing the Impact of Evaluation: A Univariate Time Series Technique.
PUB DATE Apr 79
NOTE 22p.; Paper presented at the Annual Meeting of the American Educational Research Association (63rd, San Francisco, California, April 8-12, 1979)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Decision Making; *Evaluation Methods; Goodness of Fit; *Information Utilization; *Policy Formation; Political Influences; Program Content; Program Design; Program Effectiveness; *Program Evaluation; Research Methodology; Sampling; Tests of Significance; Trend Analysis
IDENTIFIERS Meta Evaluation; Time Series Analysis

ABSTRACT

A statistical definition of information utilization for policy making decisions and an evaluation impact test to determine its occurrence are proposed. A univariate time series analysis is used to identify the internal trend for a given policy output variable and to control its effect. Two problems are identified in implementing an evaluation impact test: sample size and the nature of the time series significance test, which is used to make inferences about a series shift but actually measures goodness-of-fit of the forecasts. Other studies, involving utilization rather than quality as the dependent variable, are cited. Two general classes of evaluation characteristics related to utilization are suggested: political or the nature of the organization which will use the evaluation results and the state-of-the-art in evaluation. It is suggested that utilization might be increased if certain elements were included in evaluations, such as: a statement of the premises on which a program is based; specification of the program's process model; and analysis of the effectiveness of program components or alternative approaches. (MH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED174689

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ASSESSING THE IMPACT OF EVALUATION:

A UNIVARIATE TIME SERIES TECHNIQUE

Jonathan Shapiro

Department of Political Science

Michigan State University

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Jonathan Shapiro

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Prepared for delivery at the 1979 Annual Meeting of the American Educational Research Association, San Francisco, April 8 - 12.

TM009 661

ASSESSING THE IMPACT OF EVALUATION:
A UNIVARIATE TIME SERIES TECHNIQUE

Introduction

This paper grew out of a concern for the lack of a rigorous methodological procedure for assessing the impact of evaluations on policy decisions. Clearly, this is an important meta-evaluation issue since definitions of the purpose of evaluation often refer to the need to provide objective information such that rational policy decisions can be implemented. Carol Weiss, for example, observes that "Evaluation research is viewed by its partisans as a way to increase the rationality of decision making" (Evaluation Research: 2). Her own position is that "its purpose is to answer practical questions of decision makers who want to know whether to continue a program, extend it to other sites, modify it, or close it down" (Evaluating Action Programs: 6).

Paradoxically, while utilization is definitionally the ultimate goal of evaluation, its status as an object for research is low. Consequently, while detailed and comprehensive articles, monographs, and books explicating methodological issues such as sampling, research design, and measurement are abundant, the literature on utilization, both applied and theoretical, tends to be ad hoc and nonrigorous. As

Thomas Cook has bluntly observed, "The literature on utilization is in general not very impressive in either its form or its findings. . . . the quality and imaginativeness of most (but not all) utilization studies leaves something to be desired" (Cook: 14).

It is particularly problematic that there exists no objective set of criteria by which the impact of evaluation on policy can be measured. Thus, Weiss (1972), Anderson and Ball (1978), and Patton, et al., (1978) hypothesize characteristics of evaluation associated with utilization without ever indicating what constitutes utilization. This is somewhat analogous to constructing a regression equation without specifying how the dependent variables is to be measured.

In one recent meta-theoretical analysis of utilization, the authors admit, ". . . because of our inability to agree on an operational definition of utilization, we adopted an open-ended strategy in our interviewing that allowed respondents to define utilization in terms meaningful to them" (Patton: 62). Even assuming that decision makers are honest and realistic, the weakness of such a strategy is that without clear intersubjective agreement, the impact of an evaluation is indeterminate.

The intention of this paper is to propose a statistical definition of utilization and a procedure to test for its occurrence. Such a definition would permit measurement of utilization without resorting to subjective assessments. Recognizing that utilization may occur in nonobvious ways, it is suggested that the evaluation

impact test supplement rather than replace the existing (admittedly informal) research techniques.

An Operational Definition of Utilization

One heuristic technique often employed to support inferences on the impact of an evaluation is noting the increase (or decrease) in the level of certain policy output variables, such as funding or program operations, once the evaluation has been reported. (See, for example, The Annual Report on Education Programs, 1976, U.S.O.E.: 18-32). This notion can be expanded to provide an operational definition of utilization: Utilization has occurred when the level of a policy output variable exhibits a significant difference after evaluation.

For variables measured over time, a significant difference is one that exceeds some probability limit determined by the internal trend of the series. Since variables measured over time tend to exhibit constant growth, it is fallacious to infer that the change in level from a pre-evaluation to a post-evaluation time period is necessarily a function of the evaluation itself. Before a difference can be taken as significant, that part of the difference due to the internal trend must be discounted. Therefore, the evaluation impact test utilizes the techniques of univariate time series analysis to identify the internal trend for a given policy output variable and thereby control its effect.

A Procedure for Assessing the Impact of Evaluation

The basic assumption of univariate time series analysis is that the behavior of a series over time is due to a stochastic process which generates the series observations. By treating the observations as random realizations of a joint probability function, a time series can be characterized by its mean, variance, and covariance parameters. Nelson (page 19) notes that, in practice, we are more interested in the form of the mechanism generating observations than is the specific joint distribution function itself. The essential task is to infer the underlying stochastic process by examining the relationship between the current and past series observations. Specifying the functional form of the process allows us to explain current values and forecast future values from the history of the series.

The behavior of a policy output time series, denoted by y_t, can be represented by a linear stochastic model known as the autoregressive integrated moving average model of order p, d, q. The ARIMA (p, d, q) has the following form:

$$\phi(B)\ (1 - B)^d\ y_t = \Theta(B)a_t \qquad (1)$$

where: $y_t = (y_t - \bar{y}_t)$

ϕ and Θ are fixed

$$\phi(B) = (1 - \phi_1 B - \ldots - \phi_p B^p), \text{ the AR process} \qquad (2)$$

$$\Theta(B) = (1 - \Theta_1 B - \ldots - \Theta_q B^q), \text{ the MA process} \qquad (3)$$

$$a_t \sim \text{NID}(o, \sigma^2_{a_t}), \text{ the white noise process} \qquad (4)$$

and B is a backshift operator defined by the relation $B^m_{y_t} = y_{t-m}$. P and q refer to the order of the AR and MA processes respectively. P is the number of lagged y_t values contributing to the current value of y and q is the number of lagged a_t values contributing to the current value of y.

Identification of the Unverlying Stochastic Process

A time series can be characterized by the form and order of the stochastic process generating observations. Box and Jenkins (1970) present a procedure for identifying the stochastic process which utilizes the fact that each form of the ARIMA model exhibits a distinctive autocorrelation function (acf). The acf's associated with the AR(p), MA(q), and ARMA(p,q) processes can be derived from the distributional properties of the models.

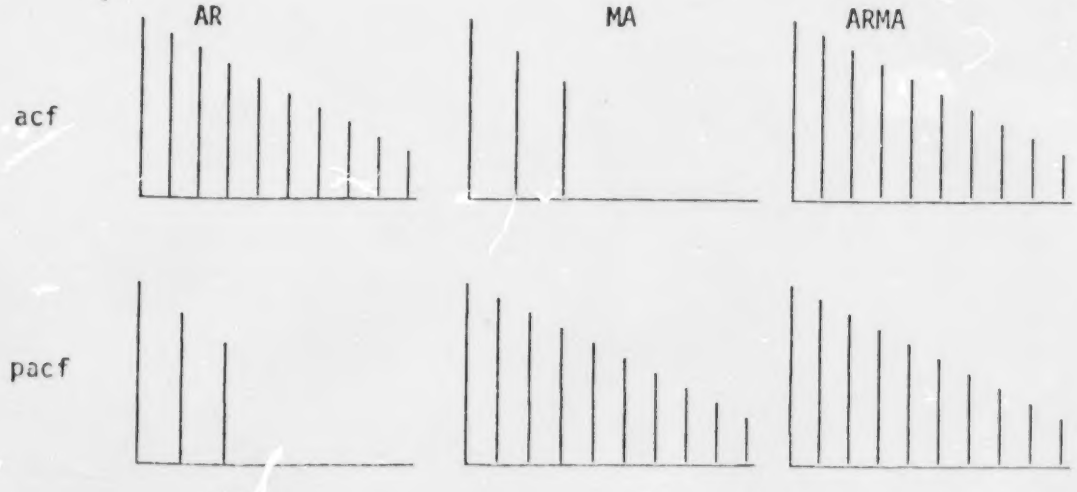
Moving average processes are indicated by a finite number of nonzero autocorrelations with an abrupt drop to zero. The acf for both the AR and ARMA processes is characterized by a series of nonzero autocorrelations exhibiting exponential decay to zero. Nonstationary series exhibit acf's with nonzero autocorrelations (close to 1) at large lags. Either a first or second difference will suffice to reduce the acf to a recognizable form.

The order of an MA(q) process is also revealed in the acf where q is equal to the number of nonzero autocorrelations. However, one consequence of an AR process is that the autocorrelations are

confounded by the influence of neighboring observations. Thus, the order of an AR or ARMA is independent of the number of nonzero autocorrelations in the acf.

The order of a pure AR can be found in the partial autocorrelation function (pacf), which essentially contains the autocorrelations controlling for the influence of neighboring observations. The order of an AR(p) is equal to the number of nonzero partial autocorrelations. The pacf for an ARMA model, however, where the autocorrelations are a function of both ϕ and θ, exhibits the same exponential decay found in the acf. Thus, the identification of the order of an ARMA (p,q) is necessarily tentative.

The theoretically derived acf's and pacf's for stationary time series models can be graphically represented as follows:



In general, in the acf, a pattern of exponential decay indicates that the process is either an AR or ARMA. An acf with a finite number of nonzero autocorrelations with an abrupt drop to zero indicates an MA process. Furthermore, the number of nonzero autocorrelations is equal to the order of the process. Thus, an MA can be completely characterized by the acf.

The pacf is used to distinguish between AR and ARMA processes. A pacf with a series of finite partial autocorrelations with an abrupt drop to zero indicates an AR process. The order of the process is equal to the number of nonzero partial autocorrelations. A pacf of exponential decline indicates an ARMA process where the order is unclear and values for p and q must be "guessed."

While this procedure seems quite straightforward, the use of sample data, contaminated by sampling error, most often results in sample correlograms that are not immediately recognizable. Identification becomes a problem of inductive inference, where the appropriate time series model must be inferred from the sample data.

Box and Jenkins propose a three-stage procedure for the identification of a time series. It involves the identification of a family of models that seem to represent the data, estimation of the data by each model to determine the model of best fit (minimum squared error) and a diagnostic checking procedure to assure that the model selected is sufficient.

Identification is the determination, based on the sample acf and pacf, of a family of time series models which seem to represent the data. One consequence of the use of sample data is that some

nonzero autocorrelations may be a function of error. To distinguish significant from nonsignificant autocorrelations, a significance test based on a result from Bartlett (Box and Jenkins: 34) is employed.

Bartlett showed that the variance for the estimated autocorrelation coefficient of a stationary normal process is approximately

$$\text{Var}\,[r_k] \simeq \frac{1}{N} \sum_{v=-\infty}^{\infty} [\rho_v^2 + \rho_{v+k}\rho_{v-k} - 4\rho_k\rho_v\rho_{v-k} + 2\rho_v^2\rho_k^2] \quad (5)$$

At the 95 percent level, the test of significance requires

$$\frac{r_k}{\text{S.E.}} \geq 2 \quad (6)$$

Often, the competing models selected in the identification stage differ from one another by a single parameter. For example, a particular acf may suggest either an AR(1) or AR(2) or a simple AR or MA versus an ARMA model. Because additional parameters use up degrees of freedom, the decision to accept a more elaborate model must be based on the marginal increase in explanation provided by the extra parameters.

Each of the models selected in the identification stage is fit to the data. To obtain the best fit, parameter values which minimize the sum of squared error are required.

The ARIMA model (equation 1) can be expressed as

$$\theta^{-1}(B)\,\phi(B)(1-B)^d\,y_t = a_t \quad (7)$$

The objective is to find a vector of AR and MA parameters such that

$$S(\phi, \theta) = \Sigma a_t^2 \tag{8}$$

is a minimum.

Two methodological problems must be considered before (8) can be solved. The first is the problem of initialization of the series. If the data set begins at $t = 1$, y_1 and a_1 are a function of the unobserved values $y_0, y_{-1}, y_{-2} \cdot \cdot \cdot$ and $a_0, a_{-1}, a_{-2} \cdot \cdot \cdot$. One solution is to set the unobserved values of y and a equal to their unconditional expected values of zero. Pindyck and Rubinfeld (482) state that this works well provided the ϕ parameters are not close to one and that $t = 1, \ldots T$ is large relative to p and q.

If the unobserved values are set to zero, the minimization time bounds are $t = 1 \ldots T$. Thus, (8) becomes

$$S(\phi, \theta) = \sum_{t=1}^{T} [a_t | \phi, \theta, Y] \tag{9}$$

where Y is the vector of observations y_t, and the errors a_t are conditioned on ϕ, θ and Y.

The second methodological problem is that the MA component of the ARIMA model is inherently nonlinear. Pure AR models can be estimated by OLS since (2) is of the form

$$y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2} + \ldots + a_t \tag{10}$$

which can be estimated as a simple linear regression.

However, since (3) indicates that for an MA, y_t is a function of the product of two polynomials, estimation must be by a nonlinear procedure. Several canned programs are available for nonlinear estimations. They are all iterative search techniques, which search a grid delimited by the stationarity conditions for the nonlinear maximum likelihood estimates. (An excellent presentation of one such technique, known as linear approximation, can be found in Pindyck and Rubinfeld, pages 482-489.)

After selecting the relatively most efficient model, the diagnostic check indicates whether the model is sufficient. If the model correctly specifies the functional form of the underlying stochastic process, the acf for the a_t's should be white noise.

A test of the residuals due to Box and Pierce (Box and Jenkins: 290 f.f.) examines the entire set of autocorrelations simultaneously. The statistic Q where

$$Q = t \sum_{j=1}^{k} r_j^2 \qquad (11)$$

is distributed as $X^2_{(k-p-q)}$df.

If Q is less than the critical X^2 value, the model is accepted as a sufficient representation of the underlying dynamic. It can be used to forecast future y_t values.

Univariate Time Series Forecasts

Once a model has been estimated and shown to have uncorrelated residuals, it can be used to generate forecasts. A forecast of y_t for period $T + \ell$, $\ell \geq 1$, is denoted by $\hat{y}_t^{(\ell)}$ and is called the origin-T forecast for lead time ℓ or the ℓ-step ahead forecast.

The most precise forecasts will be those containing little error. Because the forecast error is a random variable, minimization of the expected value of the error will yield the minimum mean square forecast error which is given by

$$\min \left(\hat{y}_t^{(\ell)} - y_{t+\ell}\right)^2 \rightarrow E\left[y_{t+\ell} \mid y_{t+\ell-1}, y_{t+\ell-2} \cdots y_1\right] \quad (12)$$

and the forecast $\hat{y}_t^{(\ell)}$ is simply the expected value of $y_{t+\ell}$ given the past observations (Pindyck and Rubinfeld: 498). The ℓ-step ahead forecast is computed recurvisely by computing $\hat{y}_t^{(1)}$, using $\hat{y}_t^{(1)}$ to compute $\hat{y}_t^{(2)}$ and so forth until $y_t^{(\ell)}$ is arrived at. Thus, $\hat{y}_t^{(\ell)}$ is given by

$$\hat{y}_t^{(\ell)} = \phi_1 y_t^{(\ell-1)} + \cdots + \phi_\ell y_t + \cdots + \phi_p y_{t-p-\ell} - \theta_1 \hat{a}_t - \cdots - \theta_q \hat{a}_{t-q+\ell} \quad (13)$$

The forecast error is the difference between the actual and forecasted values of $y_{t+\ell}$. It can be used to test the impact of evaluation on policy decisions.

The Analysis of Forecast Error as A Test of Evaluation Impact

Box and Tiao (1977) have devised a test that indicates if a time series model which adequately represented a series up to a certain point in time is still valid after an intervention occurs. They show (page 3) that since a_t is normally distributed, for $\hat{y}_t^{(\ell)}$ $\ell = 1 \ldots m$, a statistic Q can be defined where

$$Q = \sigma^{-2} \sum_{\ell=1}^{m} a_\ell^2 \sim \chi_m^2 \qquad (14)$$

Thus, the standardized sum of squares of the ℓ-step ahead forecast is distributed as χ^2 with m degrees of freedom. To ascertain whether or not the time series model is appropriate for the observations $y_1 \ldots y_m$, the standardized sum of squares can be compared to the critical value from a χ^2 table. Furthermore, Box and Tiao (page 2) note that this is equivalent to the appropriate test applied to all the lead ℓ forecast errors $e\ell$, $\ell = 1 \ldots m$.

The results due to Box and Tiao suggest the following as a test of the impact of evaluation on policy. Assume an evaluation is reported at time T + 1. Let y_t be a policy output variable such as budget appropriations or number of centers in operation over time. For y_t, $t = 1 \ldots T$, identify $\phi(B)\,(1-B)^d y_t = \Theta(B)a_t$ by identification, estimation and diagnostic checking.

Using the model, generate forecasts on the policy series for a number of post evaluation time periods. By comparing the forecasted to the actual values, the standardized sum of squares of the

forecast error can be obtained. If the Q statistic exceeds the critical X^2 value, the model is no longer appropriate and it must be the case that the post-evaluation observations are being generated by a different process than the pre-evaluation observations.

Even if the Q-test indicates that a shift in the policy output series has occurred, the shift may not necessarily be the result of the evaluation. Rival explanatory factors, the threats to internal validity, must be controlled before influence can be attributed to the evaluation itself. A quasi-experimental technique can be employed using a similar program (or programs) which has not been evaluated as the comparison. The Q-test is applied to the policy series for the comparison program resulting in the following familiar quasi-experimental design

$$\begin{array}{cccccccc} O_1 & O_2 & O_3 & X & O_4 & O_5 & O_6 \\ - & - & - & - & - & - & - \\ O_7 & O_8 & O_9 & & O_{10} & O_{11} & O_{12} \end{array}$$

where X is the evaluation. If programs similar to the one under investigation do not evidence change in the post-evaluation policy series, it can be inferred that the evaluation impacted upon the policy decision.

Two Problems of Implementation

Two problems arise in the implementation of the evaluation impact test. The first concerns sample size. Since time series analysis examines the covariance and correlation between observations,

it must rely on sample estimates of variance. It is well known that the sample variance, as an estimator of the population variance, is biased in small samples but asymptotically unbiased and consistent (Kmenta: 165). This implies that for the correlogram to be a good approximation of the acf, a large sample is required.

Furthermore, recall that one of the conditions Pindyck and Rubinfeld specified for setting unobserved values of y_t and a_t equal to zero was that t be large relative to p and q. Therefore, initializing a series for nonlinear estimation also requires a large sample. Box (page 33) has set the minimum number of observations at 50 while others have placed the minimum size from 30 to over 100. Considering the relatively short existence of most social action programs, yearly observations, while the most logical unit of analysis, will not produce sufficiently large samples. For the present, monthly or quarterly data will have to be analyzed.

The second problem concerns the nature of the time series significance test. While the test is used to make inferences about the shift in a series, it actually measures the goodness-of-fit of the forecasts. The inference concerning the shift in a series is made because it is assumed that the time series model adequately forecasted prior to the evaluation (based on the acf of the a_t's). Conceivably, if the model did not produce good forecasts prior to the evaluation, an inflated Q for the post-evaluation forecasts may be a function of the initial poor fit.

To avoid this situation, it is suggested that the model be evaluated in a manner described by Ostrom (1978). This involves the

generation of a series of ex post forecasts of a set of pre-evaluation observations excluded from the initial identification of the model. For example, given a series of y_t's, t = 1, . . . T, we could divide the sample into a set of observations t = 1 . . . t - 5, and a set of the last five observations. The initial model would be identified from the subset of y_t's, t = 1 . . . t - 5. A series of ex post forecasts, $\hat{y}^{(t-4)}$ . . . $\hat{y}^{(t)}$ can then be generated and the resulting forecast errors can be subjected to a Q test. If the test indicates an insignificant sum of squares, the model is generating forecasts of adequate fit. The excluded observations can then be re-entered into the sample to identify a model based on all the observations which can then be used to generate the post-evaluation forecasts.

If the Q-test for the ex poste forecasts is insignificant but the Q for the ex ante forecasts is significant, then the inflated Q for the post-evaluation series would indicate a shift in the level of that series.

A Meta-Theoretical Application of the Evaluation Impact Test

One popular form of theoretical analysis in evaluation is meta-evaluation or "evaluation of evaluations." Generally, such analyses examine characteristics of evaluation that relate to the quality of the data generated by evaluations. The most comprehensive analysis, due to Bernstein and Freeman (1975), examined, by regression, the relationship between 12 evaluation characteristics and the quality

of the evaluation results. This was done to make a priori predictions about the probability of obtaining high quality results given the particular circumstances under which an evaluation was initiated and implemented.

A call for an analysis of this sort with utilization rather than quality as the dependent variable has been sounded by Carol Weiss (Evaluating Action Program: 318). Weiss suggests two general classes of evaluation characteristics which may be related to utilization:

1. The (nature) of the organizational systems that are expected to use evaluation results and
2. The current state of evaluation practice

The first set of characteristics may be termed "political." Most organizations (including Congress) make decisions based on a set of information sources of which evaluation is but one. Weiss hypothesizes "Use of evaluations appears to be easiest when implementation implies only moderate alterations in procedures, staff development or costs, and where few interests are threatened" (page 320). Thus, one set of characteristics involves the degree of politicization of the results. The probability of utilization may be influenced by factors such as (1) the political salience of a program, (2) the degree of political entrenchment of a program, or (3) the degree of political support for a program.

The second set of characteristics related to utilization concerns the current state-of-the-art in evaluation. Along with methodological factors, Weiss suggests that issues such as the level of analysis may be relevant. In particular, she argues that the

informational needs of decision makers are best served by evaluations that focus on particular aspects of program structure rather than the global extent of change. Therefore, she states that utilization might be increased if evaluations include elements such as:

1. The explication of the theoretical premises underlying the program and direction of the evaluation to analyses of these premises.

2. Specification of the "process model" of the program, the presumed sequence of linkages that lead from program input to outcome, and the tracking of the processes through which results are supposed to be obtained.

3. Analysis of the effectiveness of components of the program, or alternative approaches, rather than all or nothing, go or no-go assessment of the total program (page 323).

Thus, the second set of characteristics related to utilization might include (1) methodological factors such as sampling techniques, (2) the scope of analysis, and (3) the inference generating procedure.

Taken together, the two classes of characteristics described by Weiss constitute, in a fairly explicit manner, the types of independent variables that may correlate with utilization. However, she does not indicate how the dependent variable, utilization, can be measured. This is because there are no available quantitative measures of utilization.

It is for just such a purpose, as the dependent variable in an analysis of correlates of utilization, that the evaluation impact

test would be particularly well suited. By applying the test, utilization could be measured dichotomously, depending on whether or not a shift in the level of a policy output series occurred. A probit analysis could then be performed, regressing utilization on the set of organizational and methodological characteristics outlined by Weiss. Such an analyses would be of academic interest and also prove useful to decision makers choosing between evaluation and alternative methods of information gathering for their programs.

Conclusion

In response to the manner in which evaluation impact studies are generally conducted, it has been argued that the inclusion of a quantitative procedure, focussing on changes in policy output variables, would strengthen inferences about utilization. The test proposed in this paper is based on the techniques employed in univariate time series analysis. A meta-theoretical application of the test, as the dependent variable in a study of characteristics related to utilization, has been suggested. It is hoped that additional uses of the evaluation impact procedure will be discovered in the future.

REFERENCES

Anderson, Scarvia B., and Ball, Samuel. The Profession and Practice of Program Evaluation. San Francisco: Jossey-Bass Publishers, 1978.

Box, George E. P., and Jenkins, Gwilyn M. Time Series Analysis: Forecasting and Control. San Francisco: Holden-Day, 1976.

Box, George E. P., and Tiao, G. C. "Comparison of Forecasts and Actuality." 1977. Mimeographed.

Cook, Thomas D. "Utilization, Knowledge-Building, and Institutionalization: Three Criteria by Which Evaluation Research Can Be Evaluated." In Evaluation Studies Review Annual, Vol. 3, pp. 13-22. Edited by Thomas D. Cook and Associates. Beverly Hills: Sage Publications, Inc., 1978.

Glass, Gene V.; Wilsen, Victor L.; and Guttman, John M. Design and Analysis of Time-Series Experiments. Boulder: Colorado Associated University Press, 1975.

Kmenta, Jan. Elements of Econometrics. New York: Macmilan Publishing Co., Inc., 1971.

Nelson, Charles R. Applied Time Series Analysis. San Francisco: Holden-Day, Inc., 1973.

Ostrom, Charles W., Jr. Time Series Analysis: Regression Techniques. Beverly Hills: Sage Publications, Inc., 1978.

Patton, Michael Q., et al. "In Search of Impact: An Analysis of the Utilization of Federal Health Evaluation Research." In Evaluation Studies Review Annual, Vol. 3, pp. 59-81. Edited by Thomas Cook and Associates. Beverly Hills: Sage Publications, 1978.

Pindyck, Robert S., and Rubinfeld, Daniel L. Econometric Models and Economic Forecasts. New York: McGraw Hill Book Company, 1976.

Weiss, Carol. "Evaluating Educational and Social Action Programs: A Treeful of Owls." In Evaluating Action Programs: Readings in Social Action and Education, pp. 3-27. Edited by Carol Weiss. Boston: Allyn and Bacon, Inc., 1972.

Weiss, Carol. Evaluation Research. Englewood Cliffs: Prentice-Hall, Inc., 1972.

Weiss, Carol. "Utilization of Evaluation: Toward Comparative Study." In Evaluating Action Programs: Readings in Social Action and Education, pp. 318-326. Edited by Carol Weiss. Boston: Allyn and Bacon, Inc., 1972.